说明书

发明名称: 基于RDMA的组播方法及分布式集合通信的节点设备

技术领域

[0001] 本公开涉及分布式集合通信，尤其涉及基于远程直接数据存取(Remote Direct Memory Access，RDMA)实现的适用于分布式集合通信的组播方案。

背景技术

[0002] 当前，随着人工智能生成内容(Artificial Intelligence Generated Content，AIGC)与千行百业的深度融合，大模型的参数数量持续增长、计算任务的规模也日趋庞大而复杂。这使得，分布式训练系统将会承载大规模的算力需求和任务，而分布式训练的通信效率将严重影响着分布式训练的性能。

[0003] 另一方面，RDMA作为当前主流的高性能传输方式，在分布式训练系统的通信中已经大规模应用。例如，在人工智能(Artificial Intelligence，AI)或者高性能计算(High Performance Computing，HPC)应用进行诸如allreduce、allgather等的分布式集合通信时，通常会用到RDMA这种高效的通信方式。但分布式系统中的集合通信，基本上都是基于传输层的点对点单播通信来实现多对多、一对多等的通信。具言之，应用层调用一次集合通信是由传输层的多次点对点的通信来完成。这是因为，在分布式训练中，通信初始化时，从传输层就开始进行点对点的建连，并使得后续的集合通信会基于传输层的点对点连接来进行。这种基于RDMA的点到点单播通信实现的分布式集合通信方式可能会引起较多的系统开销，还可能会造成带宽资源的浪费。

[0004] 发明内容

[0005] 本公开提供一种基于RDMA实现适用于分布式集合通信的组播方案，其可以为大规模数据中心内部的多算力协同提供更强大的支撑。

[0006] 在第一方面，本公开提供了一种基于RDMA的组播方法，应用于分布式集合通信中的源端，所述方法包括：基于第一UD类型QP发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2

的整数，并且每个目的端在确定所述组播报文携带的所述第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当基于第二UD类型QP接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[0007] 在第二方面，本公开提供了一种基于RDMA的组播方法，应用于分布式集合通信中的N个目的端中的一个所述目的端，N为大于等于2的整数，所述方法包括：基于第一UD类型QP接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；在确定所述组播报文携带的所述第一组播ID与所述目的端自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当要向所述源端回复对所述组播报文的处理结果时，基于第二UD类型QP向所述源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及所述第二组播ID，以使得所述源端在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[0008] 在第三方面，本公开提供了一种分布式集合通信的节点设备，作为所述分布式集合通信中的源端，包括处理器和支持RDMA的网卡。其中，所述处理器被配置为： 基于第一UD类型QP，经由所述网卡发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播数据报文携带的所述第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当基于第二UD类型QP经由所述网卡接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[0009] 在第四方面，本公开提供了一种分布式集合通信的节点设备，作为所述分布式集合通信中的N个目的端之一，包括处理器和支持RDMA的网卡。其中，所述处理器被配置为：基于第一UD类型QP，经由所述网卡接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；在确定所述组播报文携带的所述第一组播ID与所述目的端自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当要向所述源端回复对所述组播报文的处理结果时，基于第二UD类型QP，经由所述网卡向所述源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及所述第二组播ID，以使得所述源端在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[0010] 本公开采用的上述至少一个技术方案能够达到以下有益效果：

[0011] 通过在分布式集合通信中的源端和所有目的端各自创建支持RDMA的第一UD类型QP和第二UD类型QP，一方面，源端可基于第一UD类型QP利用网络交换机的组播功能向多个目的端的网络层发送组播报文，与常规的基于RDMA的点对点通信相比，避免了相同数据的重复发送，不仅节省了源端到网络交换机的带宽资源，还能有效降低传输时延；另一方面，每个目的端可基于第二UD类型QP向源端发送单播报文，包括向源端反馈收包情况，从而可有效提高所述分布式集合通信的可靠性。

[0012] 特别地，在源端和所有目的端进入分布式集合通信的数据组播流程之前，由源端基于第一UD类型QP以组播的方式通告待传输数据的数据信息，而每个目的端基于第二UD类型QP以单播的方式回复通告确收信息，使得源端和所有目的端能够在数据组播流程中实时更新、维护各自的数据包收发记录，并且源端能够在针对特定数据块的组播流程结束后基于自身的发包记录和每个目的端反馈的收包记录确认所有目的端是否均正确接收了所发送的数据包，并视情况触发对未被所有目的端正确接收的数据包的重传。这样，与接收端实时回复收包确认的方式相比，不仅可显著减少确收报文的数量，还避免了因实时反馈丢包信息导致的等待时长，从而有效提升分布式集合通信的效率。

附图说明

[0013] 此处所说明的附图用来提供对本公开的进一步理解，构成本公开的一部分。本公开的示意性实施例及其说明用于解释本公开，不构成对本公开的不当限定。

[0014] 图1为根据本公开实施例的一种用于实现分布式集合通信中一对多组播的方式的示意图。

[0015] 图2为根据本公开实施例的另一种用于实现分布式集合通信中一对多组播的方式的示意图。

[0016] 图3A和图3B为根据本公开实施例的利用统一通信接口(Unified Communication X，UCX)进行基于RDMA的分布式训练的节点架构示意图。

[0017] 图4为根据本公开实施例的又一种用于实现分布式集合通信中一对多组播的方式的示意图。

[0018] 图5A为根据本公开实施例的一种用于建立适用于分布式集合通信的基于RDMA的组播连接的方法的流程示意图。

[0019] 图5B和图5B(续)为根据本公开实施例的一种适用于分布式集合通信的基于RDMA的组播方法的流程示意图。

[0020] 图6为根据本公开实施例建立的源端的发包状态记录的示意图。

[0021] 图7为根据本公开实施例建立的目的端的收包状态记录的示意图。

[0022] 图8为根据本公开实施例的一种基于RDMA的组播方法的流程示意图，其应用于分布式集合通信中的源端。

[0023] 图9为根据本公开实施例的一种基于RDMA的组播方法的流程示意图，其应用于分布式集合通信中的N个目的端之一。

[0024] 图10为根据本公开实施例的一种分布式集合通信的节点设备的功能模块图。

具体实施方式

[0025] 为使本公开的目的、技术方案和优点更加清楚，下面将结合本公开具体实施例及相应的附图对本公开技术方案进行清楚、完整地描述。所描述的实施例仅是本公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，本领域

普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

[0026] 以下结合附图，详细说明本公开各实施例提供的技术方案。

[0027] 在当前的分布式系统中，除了点对点通信外，集合通信更多为一对多、多对多的情况。由于当前主流的一些传输层协议仅支持点到点的传输方式，因此，集合通信中一对多、多对多的数据传输也大多基于点到点的数据传输方式来实现，即多次调用点对点的传输来完成一对多、多对多的数据转发。

[0028] 以一对多广播(Broadcast，一个数据发送者，多个数据接收者)为例，需要在集群内把一个节点上的数据广播到其他多个节点上。当前，一种用于实现一对多广播的方式可如图1所示，源节点(以下也可称为源端)110将同一份消息依次经由网络交换机120转发给不同的目的节点(以下也可称为目的端)130−1、130−2、…130−N。其中，尽管图示N大于2，本领域技术人员应能理解，N也可以等于2。换言之，N表示目的节点的数量，只要N为大于或等于2的整数，即可达成所谓的一对多广播。假若数据量为M，带宽为B，n为节点数，该方式实现一对多广播的总耗时可如下计算：

$$T_{Bcast} = (\mathrm{n}-1) * \left(t_s + \frac{M}{B}\right) = (n-1) * \frac{M}{B} + (n-1) * t_s$$

[0029] 其中，t_s表示启动单个传输的启动时间，包括网络层协议栈处理、CPU中断应用层协议栈处理、数据转发到网口总耗时。

[0030] 显然，该方式会引入与通信规模线性相关的额外开销，造成时延的增大。并且，当源端110到网络交换机120的带宽较大时，该方式容易造成带宽资源的浪费；而当分布式训练的数据量较大时，该方式又容易形成带宽瓶颈链路。

[0031] 目前，另一种用于实现分布式集合通信中一对多广播的主流方式可如图2所示。在源节点210和多个目的节点230−1、230−2、…230−N的通信过程中，采用目的节点230−1、230−2、…230−N−1作为协助进行中间数据转发。具体为，源端210将要发送的消息拆分成多个数据块，每次发送一个数据块，第一个目的节点230−1收到完整的数据块后将该数据块转发给下一个目的节点230−2。依此类推，使得该数据块被依次转发到最后一个目的节点230−N。由于源端210可在该

数据块在目的节点230−1、230−2、…230−N之间传递的期间向第一个目的节点发送下一个数据块，这种方式可以实现数据块的流水线并行发送，有效提高了分布式系统的节点间通信效率。

[0032] 具言之，假如数据量为M，带宽为B，n为节点数，源端将M数据分为s块，上述方式实现一对多广播的总耗时可如下计算：

$$T_{Bcast} = \frac{(s+n-2)M}{s*B} + (s+n-2)*t_s$$
$$= \frac{(s+n-2)}{s} * \frac{M}{B} + (s+n-2)*t_s$$

[0033] 其中，t_s表示启动单个传输的启动时间，包括网络层协议栈处理、CPU中断应用层协议栈处理、数据转发到网口总耗时。

[0034] 上述实现方式虽然在一定程度上提升了通信效率，减少了带宽瓶颈，但仍然存在与通信规模成线性相关的额外开销。并且，协助进行中间数据转发的目的节点230−1、230−2、…230−N−1在收到报文后需要进行报文的复制和发送，也会增加额外其上CPU中断和处理的时延。

[0035] 有鉴于此，本发明旨在提出一种适用于分布式集合通信的基于RDMA的组播方法，通过利用网络交换机完成组播报文的复制分发，以网络层组播替代应用层组播，来避免相同数据的重复发送，可有效节省源端到网络交换节点的带宽，并因减少了中间协助转发数据的处理时间，还可有效降低传输时延。

[0036] 统一通信接口(Unified Communication X，UCX)旨在提供一个统一的抽象通信接口，以对基于RDMA的通信进行优化，并支持各种应用的需求。在当前分布式系统的集合通信中，为了优化基于RDMA的数据传输，大多数的通信库(如NVIDIA集合通信库(NVIDIA Collective Communications Library，NCCL)、开源消息传递接口(Open Message Passing Interface，OpenMPI)等)均已兼容UCX，如图3A和图3B所示。其中，UCX是基于端点(endpoint)来进行节点间通信的，即源端使用UCX时需要创建一个端点、并与目的端所创建的端点进行连接。并且，每个端点用相应的端点ID来唯一标识。此外，在AI或HPC应用中，例如进行一个AI分布式训练时，通常会启动多个进程，且如NCCL、OMPI之类的通信库会给每个进程分配一个排序值rank，以能够唯一标识相应的进程。

[0037] 因此，本发明主要基于UCX和RDMA来实现适用于分布式集合通信的网络层组播(以下也可简称为网络RDMA组播)，以利用网络交换机的组播功能实现分布式集合通信。具言之，当源端的分布式训练进程要调用broadcast通信原语广播数据时，可基于底层UCX建立组播连接(以下也可称为进行组播建连)，然后利用所建立的组播连接发送组播数据报文，以利用网络交换机的组播功能将该组播数据报文转发给所有的目的端。相应地，当目的端收到该组播数据报文时，可回复源端ACK/NACK。这样，如果出现丢包，源端会进行重传，从而可以有效确保组播的可靠性。

[0038] 上述网络RDMA组播通信的框架可如图4所示。其中，源端410与多个目的端430-1、430-2、…430-N之间利用UCX建立了基于RDMA的组播连接，从而能够利用网络交换机420的组播功能实现适用于分布式集合通信的基于RDMA的组播。

[0039] 若数据量为M，最大传输单元(Maximum Transmission Unit，MTU)为U，带宽为B，n为节点数，上述方案实现一对多广播的耗时可如下计算：

$$T_{Bcast} = \frac{M}{B} + \frac{M}{U} * (t_s + T_0)$$

[0040] 其中，T_0为组播发送一次数据引起的额外时延，组播需要在数据的有效载荷(payload)中额外增加一些头信息，因此数据解析、封装都相比前述两种方式额外引入了部分时延。t_s表示启动单个传输的启动时间，包括网络层协议栈处理、CPU中断应用层协议栈处理、数据转发到网口总耗时。

[0041] 由于UCX是基于点对点的通信框架，为实现可靠的网络RDMA组播方案，根据本发明的实施例，首先修改UCX通信框架以实现一对多的组播建连。

[0042] 1、基于UCX的适用于分布式集合通信的组播建连

[0043] 首先，参与分布式集合通信的源节点和目的节点需要建立基于RDMA(例如，符合RoCEV2协议的RDMA)的组播连接。

[0044] 具言之，由于RDMA的不可靠数据报(Unreliable Datagram，UD)类型的队列对(Queue Pair，QP)无法保证目的端会收到数据包，也不能保证数据包的接收顺序。因此，为了能够基于RDMA实现适用于分布式集合通信的可靠组播连接，

在根据本发明实施例的技术方案中，在源端和所有的目的端创建两个UD类型QP，其中：第一UD类型QP用于建立源端与所有目的端之间的组播连接，即使源端和所有目的端构建为一个多播组，以使得源端发送的组播报文可利用网络交换机的组播功能被转发至各个目的端，从而实现分布式集合通信；而第二UD类型QP用于建立源端与各个目的端之间的单播映射(以下也可称为虚拟单播链路)，以使得各个目的端可向源端发送单播报文，包括回复ACK/NACK报文，从而可有效保证分布式集合通信的可靠性。

[0045] 其次，还需要对UCX的点对点通信框架进行修改，即突破UCX的点对点通信框架的限制对单播链路进行扩展，以实现一种基于UCX的点对多点的建连方案，从而保证目的端能正确识别源端发出的组播数据报文、并且目的端能可靠地单播回复ACK/NACK给源端。

[0046] 该建连方案可如图5A所示。如图5A所示，该建连方案可大体划分为两个阶段，即建立组播连接和建立虚拟单播链路。

[0047] 其中，如图5A中虚线之上部分所示，建立组播连接的阶段可具体包括：

[0048] ①源端510创建两个UD类型QP，和②各目的端530-1、…530-N也创建两个UD类型QP。其中，第一UD类型QP用于构建、加入多播组，第二UD类型QP用于建立虚拟单播链路。

[0049] 需注意的是，尽管图5A将步骤①示出为在步骤②之前执行，实际上，这两个步骤的执行顺序不做限定，即步骤①还可与步骤②同时执行、或者在步骤②之后执行。

[0050] ③源端510生成唯一的组播互联网协议(Internet Protocol，IP)地址。

[0051] 相应地，所有目的端530-1、…530-N可通过分布式集合通信中原有的点对点传输方案获取到该组播IP地址。

[0052] ④源端510和所有目的端530-1、…530-N根据该组播IP地址按照相同的规则生成相同的3字节的组播ID(groupid)，用以标识所属的组播，也即用以标识源端和所有目的端利用前述第一UD类型QP创建的组播组。

[0053] 就此，源端510和所有目的端530-1、…530-N之间成功建立了组播连接。具言之，若源端510发送组播数据报文时携带此组播ID groupid，则目的端530-1、

…530-N可在收到该报文后通过根据自身保存的组播ID groupid与该报文中携带的组播ID groupid进行匹配来正确识别、处理该组播数据报文。其中，组播ID groupid满足最高比特(bit)位为1(如0x800001)即可。

[0054] 此外，如图5A中虚线之下部分所示，建立虚拟单播链路的阶段可具体包括：

[0055] ⑤源端510随机生成一个3字节的虚拟端点ID v_epid_0_0，而所有目的端根据自身所属的排序值rank按照一定规则生成不同的3字节的虚拟端点ID v_epid_1_0、….v_epid_N_0，用以标识各节点收发的报文。如前所述，UCX是基于端点(endpoint)来进行节点间通信的，且利用端点ID来唯一标识相应的端点。在根据本发明的实施例中，为了基于UCX实现点对多点的组播建连，在UCX的端点ID基础上扩展出虚拟端点ID，主要用于识别收发报文的各节点。其中，源端510生成的虚拟端点ID v_epid_0_0和各目的端生成的虚拟端点ID v_epid_1_0、….v_epid_N_0均满足最高bit位为0(如0x000001)即可。

[0056] ⑥源端510发起连接请求、即生成CREQ报文，并通过组播方式发送该CREQ报文。其中，该CREQ报文携带源端自身的虚拟端点ID v_epid_0_0、组播ID以及合格供应商网络(qualified provider network，QPN)等信息。

[0057] ⑦所有目的端530-1、…530-N收到来自源端510的CREQ报文之后，通过单播方式给源端回复连接请求响应、即CREP报文。其中，此CREP报文中携带目的端自己生成的虚拟端点ID v_epid_1_0、….v_epid_N_0、组播ID以及QPN等信息，并且此报文中会携带确收(ACK)标志位，以确保源端510收到该CREP报文后能够向对应的目的端回复ACK报文。

[0058] ⑧当源端510陆续收到来自各个目的端530-1、…530-N的CREP报文后，保存这些CREP报文中携带的信息。需要说明的是，源端510接收来自各个目的端530-1、…530-N的CREP报文的次序不确定，且不一定与各个目的端530-1、…530-N接收来自源端510的CREQ报文的顺序相同。因此，源端510会对接收到的CREP报文进行计数，直到收到所有目的端530-1、…530-N回复的携带ACK信息的CREP报文为止。并且，在确认已接收到所有目的端530-1、…530-N回复的携带ACK信息的CREP报文后，源端510可根据所接收到的CREP报文中的ACK信息，通过组播方式回复ACK报文给所有目的端530-1、…530-N。

[0059] 特别地，若所接收到的CREP报文携带NACK、而非ACK信息，则源端510会重新通过组播方式发送所述CREQ报文。在这种情况下，此前已回复携带ACK信息的CREP报文的目的端会对重复收到的CREQ报文进行忽略处理，而此前所回复的CREP报文携带NACK信息的目的端会对重复收到的CREQ报文进行处理，并在处理正确之后回复携带ACK信息的CREP报文给源端510。

[0060] 通过以上处理，成功建立了用于各目的端530-1、…530-N单播回复报文(可携带ACK/NACK信息以进行收包确认和/或丢包报告)给源端510的虚拟单播链路，具体包括基于虚拟端点ID建立的单播映射{v_epid_0_0,v_epid_1_0}、…{v_epid_0_0,v_epid_N_0}。

[0061] 到此，源端可基于前述第一UD类型QP向多个目的端发送组播报文，而每个目的端可基于前述第二UD类型QP向源端发送单播报文。其中，所述组播报文携带源端的虚拟端点ID以及所述组播ID，以使得所述组播报文能够经由所述网络交换机的组播功能转发至每个目的端，且每个目的端通过将所述组播报文携带的所述组播ID与自身的组播ID进行比较来识别、处理所述组播报文；所述单播报文携带所述组播ID以及所述目的端自身的虚拟端点ID，以使得所述源端能够正确识别、处理所述单播报文。换言之，通过以上操作，基于UCX的组播建连完成。

[0062] 2、适用于分布式集合通信的网络RDMA组播实现

[0063] 为了确保网络RDMA组播传输分布式集合通信数据的可靠性，需要设计一种保障机制。一般来说，可要求每个数据包的接收端都回复确收报文，以确保报文的正确接收。但这种方式显然会浪费网络带宽。同时，由于网络的不稳定性，目的端回复的确收报文存在丢失的可能，并进一步造成正确接收的数据包被重传，而这显然也会导致网络资源的浪费，进而影响分布式集合通信的效率。

[0064] 因此，根据本发明的实施例，在前期基于UCX建立了适用于分布式集合通信的组播连接的基础上，设计了一种适用于分布式集合通信的网络RDMA组播方式。

[0065] 如图5B和图5B(续)所示，该网络RDMA组播方法可大体划分为三个阶段，即信息通告、数据组播和收包确认。其中，信息通告主要为后续两个阶段提供初始

化准备条件，包括在源端和所有目的端之间同步待传输数据的数据信息，并建立相应的发包状态记录和收包状态记录；数据组播主要用于源端以组播的方式发送待传输数据包含的数据包，源端和所有目的端会在该过程中实时更新自身的发包状态记录和收包状态记录；而收包确认主要用于在结束针对特定数据块的数据包传输之后，基于相应的发包状态记录和收包状态记录进行数据包接收情况确认。

[0066] 如图5B所示，信息通告阶段可具体包括以下步骤①至步骤④。

[0067] ①分布式集合通信中的源端510根据数据量大小，按照一定阈值将待传输的数据共切分成K个数据块。其中，由于利用UD类型QP传输数据的不可靠性，推荐以MTU作为支持的最大消息。因此，每个数据块可按照MTU大小发送数据，共发送[数据块的数据量/MTU]个数据包。其中，⌈⌉表示向上取整运算符，假设[数据块的数据量/MTU]等于s，则每个数据块被进一步划分为s个数据包。

[0068] 分布式集合通信中的源端510根据每个数据块中包含的数据包的个数s，按照s个数据包的顺序建立发包状态记录。其中，发包状态记录可采用以数据包的序列号作为位置索引在相应地位置记录数据包的状态标识的发包状态图send_pkt_map，具体可如图6所示。其中，N表示分布式集合通信中目的端的个数；二值化的状态标识pkt1、pkt2、…、pkti、…、pkts分别标识s个数据包的状态，可取值0或1。例如，若状态标识pkti取值0，表示第i个数据包尚未发送；而若状态标识pkti取值1，则表示第i个数据包已发送。I表示数据包的序列号，为大于0小于s的整数。

[0069] ②当准备好通过网络RDMA组播的方式发送数据报文时，源端510会先构造一个包含数据信息的组播信息通告报文，并在该组播信息通告报文的data起始位置携带源端的虚拟端点ID v_epid_0_0、组播ID groupid以及信息通告标识tag0。其中，所述数据信息可至少包括待传输数据的数据量M、被划分为K个数据块、每个数据块被进一步划分为s个数据包、以及各数据包的序列号等信息。

[0070] 然后，源端510基于在前面创建的第一UD类型QP发送该组播信息通告报文，并启动第一定时器。同时，源端510将发包状态记录中s个数据包的状态标识

pkt1、pkt2…pkts均初始化置为0。后续，每当源端510发送对应序列号的数据包时，会将对应的状态标识改为1。

[0071] ③当目的端530-1、…530-N收到来自源端510的组播信息通告报文时，可通过该组播信息通告报文中携带的虚拟端点ID v_epid_0_0、组播ID groupid以及信息通告标识tag0确定当前组播组的源端510需要发送的数据信息，并建立相应的收包状态记录。其中，收包状态记录可采用以数据包的序列号作为位置索引并在相应地位置记录数据包的状态标识的收包状态图recv_pkt_map，具体可如图7所示。其中，收包状态记录中s个数据包的状态标识pkt1、pkt2…pkts均被初始化置为0。这样，后续每当目的端收到对应序列号的数据包时，可将对应的状态标识改为1。

[0072] 同时，每个目的端530-1、…530-N利用此前创建的第二UD类型QP单播回复源端510信息通告确收报文(以下也可称为信息通告ACK报文)，以使得源端能够确认该组播信息通告报文被该目的端正确收到。特别地，目的端530-1、…530-N回复的信息通告ACK报文中携带目的端自身的虚拟端点ID v_epid_1_0、….v_epid_N_0和组播ID groupid，以便源端510进行收包确认。

[0073] ④如果在第一定时器的计时范围(以下也可称为第一预定时长)内收集到所有目的端530-1、…530-N回复的信息通告ACK报文，则源端510开始启动数据报文的组播发送。而如果在第一定时器计时结束时(即从发送组播信息通告报文起已过去了所述第一预定时长)仍未收集全所有目的端回复的信息通告ACK报文(包括未收到任何目的端回复的信息通告ACK报文)，则源端510会重新发送该组播信息通告报文、并再次启动第一定时器，直到源端510收集到所有目的端回复的信息通告ACK报文为止。

[0074] 由此可见，收集到所有目的端530-1、…530-N回复的信息通告ACK报文，是源端510开启数据组播发送的前提条件。特别地，为了避免因可能的信息通告ACK报文丢失而导致无休止的收集等待，可利用第一定时器来约束等待时间。具言之，若第一定时器计时结束时仍未收集到所有目的端530-1、…530-N回复的信息通告ACK报文，则源端510可推断至少一个目的端未能够正常接收该组播

信息通告报文、和/或至少一个目的端回复的信息通告ACK报文被丢失，并重新发送该组播信息通告报文，同时重启第一定时器。

[0075] 如图5B和图5B(续)所示，数据组播流程主要包括以下步骤⑤和步骤⑥。

[0076] ⑤源端510开始启动数据报文的组播发送流程。具言之，源端510会基于要发送的数据包构造组播数据报文，在该组播数据报文data起始位置携带源端的虚拟端点ID v_epid_0_0、组播ID groupid以及数据报文标识tag1，并利用前面创建的第一UD类型QP发送该组播数据报文，以确保该组播数据报文可经过网络中交换机的组播功能转发到各目的端530-1、…530-N。

[0077] 同时，源端510将发包状态记录中与所发送的数据包对应的状态标识置为1。

[0078] ⑥当目的端530-1、…530-N收到该组播数据报文时，可通过其携带的虚拟端点ID v_epid_0_0、组播ID groupid以及数据报文标识tag1正确识别该报文，并根据所携带的数据包的序列号psn置位自身收包状态记录中的相应位置。

[0079] 具言之，目的端530-1、…530-N根据源端510发送的组播数据报文携带的包序列号(Packet Sequence Number，PSN)，把自身的收包状态记录中对应位置的状态标识置为1。例如，假设源端510发送多个报文的PSN从1开始依次递增至s，目的端530-1、…530-N收到的组播数据报文中携带的序列号为i，则目的端530-1、…530-N将把自身收包状态记录中对应的第i个位置的状态标识pkti取值1。

[0080] 源端510和目的端530-1、…530-N可分别重复上述步骤⑤和⑥，直至将同一数据块的所有s个数据包组播发送完毕。

[0081] 然后，如下文所述，源端510将启动针对该数据块的收包确认，并在确认所有目的端530-1、…530-N均正确接收到该数据块的所有s个数据包之后，结束该数据块的组播发送流程，并启动下一个数据块的组播发送流程。如图5B(续)所示，针对该数据块的收包确认主要包括以下步骤⑦至步骤⑨。

[0082] ⑦当源端510发送完一个数据块的所有s个数据包时，需要进行目的端的收包确认。这时，源端510会构造一个组播状态通告报文，在该组播状态通告报文的data起始位置携带源端的虚拟端点ID v_epid_0_0、组播ID groupid以及状态通告标识tag2，并启动第二定时器。

[0083] ⑧当目的端530−1、…530−N收到该组播状态通告报文时，会基于自身的收包状态记录构造状态响应报文，并利用前面创建的第二UD类型QP将该状态响应报文单播发送给源端510。特别地，该状态响应报文还携带目的端自身的虚拟端点ID v_epid_1_0、…v_epid_N_0和组播ID groupid，以便源端510进行确认。

[0084] 需要说明的是，目的端530−1、…530−N有可能在收到组播状态通告报文时还未收到该数据块的某个数据包、即尚未收全该数据块的所有s个数据包。在这种情况下，目的端530−1、…530−N将基于自身当前的收包状态记录构造状态响应报文并单播发送给源端510，以使得源端510能够准确获知目的端上的数据包接收状态，并相应触发针对目的端尚未接收到的数据包的重传。

[0085] 例如，假设目的端530−t在收到组播状态通告报文时，已收到该数据块中除第2个数据包之外的其他s−1个数据包；而剩余N−1个目的端530−1、…530−t−1、530−t+1、…530−N在收到组播状态通告报文时，均已成功收到该数据块的所有s个数据包。换言之，在收到组播状态通告报文时，目的端530−t上的收包状态记录中的状态标识pkt2取值0、而剩余的s−1个状态标识均取值1；并且，目的端530−1、…530−t−1、530−t+1、…530−N上的收包状态记录中的所有s个状态标识均取值1。在这种情况下，目的端530−1、…530−t−1、530−t、530−t+1、…530−N将经由状态响应报文，把上述数据包接收状态信息如实反馈给源端510，以使得源端510在执行下文步骤⑨中的操作之后，其上发包状态记录中第二个数据包对应的状态标识为(1，0)、而非表示所有目的端均正确接收该数据包的(0，1)。据此，源端510将确认有一个目的端没有收到第二个数据包，并将返回步骤⑤重传该第二个数据包。

[0086] ⑨源端510如果在第二定时器的计时范围(以下也可称为第二预定时长)内未收到任何目的端回复的状态响应报文，则重传该组播状态通告报文。否则，源端510会根据接收到的状态响应报文携带的虚拟端点ID v_epid_1_0、….v_epid_N_0以及收包状态记录更新发包状态记录，包括根据该状态响应报文中收包状态记录对发包状态记录进行对应位置的与运算(以下用运算符“^”表示)。

[0087] 具体为，在第二定时器的计时范围内，每次收到一个目的端回复的状态响应报文，源端510会针对根据该状态响应报文携带的收包状态记录中的每个状态标识

pkti，进行如下处理：若pkti取值为1，则将发包状态记录中的(J,pkti)更新为(J−1,pkti^1)；若pkti取值为0，则将发包状态记录中的(J,pkti)更新为(J,pkti^0)。其中，i表示数据包的序列号，为小于等于s且大于等于1的整数；J为小于等于N且大于等于1的整数，可用于表示未接收到第i个数据包的目的端的个数，也即N−J可表示对正确接收到第i个数据包的目的端计数。

[0088] 例如，假设1个源端、3个目的端(N＝3)、每个数据块被划分为5个数据包(s＝5)，则源端510将一个数据块中的5个数据包均发出去之后，发包状态记录为[(3,1),(3,1),(3,1),(3,1),(3,1)]。

[0089] 然后，假设目的端530−1收全了5个数据包，其上收包状态记录为[1,1,1,1,1]。则源端510收到目的端530−1回复的状态响应报文后，将根据此收包状态记录[1,1,1,1,1]将发包状态记录从[(3,1),(3,1),(3,1),(3,1),(3,1)]更新为[(2,1),(2,1),(2,1),(2,1),(2,1)]。

[0090] 又假设目的端530−2收到除第2个数据包之外的4个数据包，其上收包状态记录为[1,0,1,1,1]。则源端510收到目的端530−2回复的状态响应报文后，将根据此收包状态记录[1,0,1,1,1]将发包状态记录进一步从[(2,1),(2,1),(2,1),(2,1),(2,1)]更新为[(1,1),(2,0),(1,1),(1,1),(1,1)]。

[0091] 再假如目的端530−3收到除第2个数据包和第5个数据包之外的3个数据包，其上收包状态记录为[1,0,1,1,0]。则源端510收到目的端530−3回复的状态响应报文后，将根据此收包状态记录[1,0,1,1,0]将发包状态记录进一步从[(1,1),(2,0),(1,1),(1,1),(1,1)]更新为[(0,1),(2,0),(0,1),(0,1),(1,0)]。

[0092] 在第二定时器计时结束时(即从发送组播状态通告报文起过去了第二预定时长)，源端510依次检查发包状态记录中各数据包的状态标识，并在确认所有数据包的状态标识均为(0，1)时，结束该数据块的发送，并清除自身上该数据块的缓存。如果发包状态记录中存在状态标识非(0，1)的数据包，则源端510将再次执行步骤⑤中利用第一UD类型QP发送组播数据报文的操作，以重传该数据包。相应地，目的端收到重传的数据包后，若此前没有接收到该数据包，则如前所述进行接收处理，并相应更新自身的收包状态记录中的对应位置，同时基于更新后的收包状态记录构造前述状态响应报文并单播回复给源端510，以使得源端510

可相应更新发包状态记录；而若此前已正确接收该数据包，则删除重复报文即可，也不用再次回复状态响应报文给源端。

[0093] 继续以前述示例说明。源端510通过检查发包状态记录[(0,1),(2,0),(0,1),(0,1),(1,0)]可知，第1、第3、第4个数据包被所有目的端正确接收，但第2个数据包有2个目的端没有收到，第5个数据包有1个目的端没有收到，并因此触发重传第2、第5个数据包的数据报文组播。

[0094] 相应地，目的端530-2收到重传数据包的组播数据报文后，会发现第5个数据包已收到过，则忽略；而第2个数据包之前没收到，则进行接收处理，并将其上的收包状态记录从[1,0,1,1,1]更新为[1,1,1,1,1]。然后，目的端530-2将基于自身更新后的收包状态记录构造新的状态响应报文并单播回复给源端510，以使得源端510可响应于该新的状态响应报文将发包状态记录从[(0,1),(2,0),(0,1),(0,1),(1,0)]更新为[(0,1),(1,0),(0,1),(0,1),(1,0)]。需要注意的是，在源端510基于更新的收包状态记录[1,1,1,1,1]更新发包状态记录时，将仅对发包状态记录中非(0,1)的状态标识进行更新。这是因为，状态标识(0,1)表示对应的数据包已被所有目的端正确接收，从而无需更新。

[0095] 此外，目的端530-3收到重传数据包的组播数据报文后，会发现第2、第5个数据包之前都没收到过，则进行接收处理，并将其上的收包状态记录从[1,0,1,1,0]更新为[1,1,1,1,1]。然后，目的端530-3将基于自身更新后的收包状态记录构造新的状态响应报文，并单播回复给源端510，以使得源端510可响应于该新的状态响应报文将发包状态记录从[(0,1),(1,0),(0,1),(0,1),(1,0)]更新为[(0,1),(0,1),(0,1),(0,1),(0,1)]。类似地，源端510此次更新发包状态记录时也仅对其中非(0,1)的状态标识进行更新。

[0096] 至此，源端510可根据发包状态记录[(0,1),(0,1),(0,1),(0,1),(0,1)]确认属于同一数据块的5个数据包均被所有目的端正确接收了，从而结束该数据块的组播发送流程，并可以启动下一个数据块的组播发送流程了。例如，源端510可返回执行上面步骤⑤中的操作，直到步骤②中组播信息通告报文所涉及的要发送数据全部发送完毕、且被所有目的端正确接收。

[0097] 最后，结合图5B和图5B(续)可知，根据本发明实施例的适用于分布式集合通信的网络RDMA组播方案，在数据组播阶段，源端经由第一UD类型QP发送携带源端的虚拟端点ID、组播ID的组播数据报文，以利用网络交换机完成组播报文的复制分发，避免了相同数据的重复发送，不仅节省了源端到网络交换机的带宽，还有效降低了传输时延。此外，通过在信息通告阶段进行待传输数据被切分为多个数据块、且每个数据块被进一步切分为多个数据包的信息同步，并在收包确认阶段基于在数据组播阶段实时更新的发包状态记录和收包状态记录以数据块为单位进行数据包的传输确认，不仅显著减少了确收报文的数量，还避免了实时反馈丢包信息导致的等待反馈时间，从而能够有效提升分布式集合通信的效率。

[0098] 基于以上具体示例的详细介绍，根据本公开的实施例，提供了一种基于RDMA的组播方法，其可应用于分布式集合通信中的源端。如图8所示，该组播方法可包括：步骤S810，基于第一UD类型QP发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播报文携带的所述第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及步骤S820，当基于第二UD类型QP接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[0099] 可选地，如图8所示，在步骤S810之前，所述方法还包括：在源端预先创建第一UD类型QP和第二UD类型QP(步骤S810p1)；以及，在源端预先生成第一组播ID和第一虚拟端点ID(步骤S810p2)。

[0100] 可选地，步骤S810p2具体包括：随机生成第一虚拟端点ID；根据组播IP地址，按照预定的规则生成所述第一组播ID；以及向每个目的端发送所述组播IP地址，以使得每个目的端根据所述组播IP地址按照所述预定的规则生成第二组播ID。这样，源端的第一组播ID以及各目的端的第二组播ID完全相同，可用来标

识由源端和所有目的端构建的组播组内的报文。此外，源端的第一虚拟端点ID以及各目的端的第二虚拟端点ID相互不同，可用来标识在该组播组内报文的具体来源。

[0101] 可选地，参考图5B所示，在基于第一UD类型QP发送携带数据报文标识以及待发送的数据包的组播数据报文(步骤⑤)之前，所述方法还包括：基于所述第一UD类型QP发送组播信息通告报文(步骤②)，其中，所述组播信息通告报文携带有信息通告标识以及与待传输数据相关的数据信息，所述数据信息包括待传输数据的数据量M、待传输数据被划分为K个数据块、每个数据块被进一步划分为s个数据包、以及各数据包的序列号，K为大于或等于1的整数，s为大于或等于1的整数；以及，基于第二UD类型QP接收N个目的端返回的信息通告确收报文(步骤③)，其中，每个信息通告确收报文携带有表示目的端接收到所述组播信息通告报文的确收标识。

[0102] 进一步，如前所述，为了约束接收N个信息通告确收报文的等待时间，参考图5B所示，基于第二UD类型QP接收N个目的端返回的信息通告确收报文，可具体包括：在第一预定时长计时结束时判断已接收到的信息通告确收报文的个数是否小于N(步骤④)，其中，第一预定时长从发送所述组播信息通告报文时起算；若前述判断的结果为是，则重新发送所述组播信息通告报文，并重新启动所述第一预定时长的计时，以返回前述判断步骤④。

[0103] 此外，通过在各目的端上维护并更新收包状态记录，源端可以基于自身的发包记录以及各目的端上的收包状态记录，以数据块为单位进行批量的收包确认。具言之，参考图5B(续)所示，针对K个数据块中的每个数据块，在发送完该数据块包含的所有待发送数据包之后，所述方法还包括：基于第一UD类型QP发送针对该数据块的组播状态通告报文，其中，所述组播状态通告报文携带有状态通告标识(步骤⑦)；基于第二UD类型QP接收状态响应报文，其中，所述状态响应报文携带有目的端自身的收包状态记录(步骤⑧)；依据该数据块下各数据包的发送记录以及接收到的状态响应报文中的收包状态记录，确定N个目的端是否接收到该数据块下的已发送数据包(步骤⑨)。

[0104] 可选地，依据该数据块下各数据包的发送记录以及接收到的状态响应报文中的收包状态记录，确定N个目的端是否接收到该数据块下的已发送数据包，可具体包括：依据该数据块下各数据包的发送记录以及在第二预定时长内接收到的每个状态响应报文携带的收包状态记录，判断该数据块下的每个已发送的数据包是否被所有N个目的端接收，其中，第二预定时长从发送所述组播状态通告报文起算；若确定存在未被所有N个目的端接收的数据包，则基于第一UD类型QP发送携带该数据包的组播数据报文，并重新启动所述第二预定时长的计时；返回前述判断处理，直到确定该数据块包括的所有数据包均已被所述N个目的端接收。

[0105] 进一步，如前所述，为了约束接收响应状态报文的等待时间，如图5B(续)所示，若在所述第二预定时长计时结束时未接收到任一状态响应报文，所述方法还包括：重新发送所述组播状态通告报文，并重新启动所述第二预定时长的计时。

[0106] 相应地，根据本公开的一实施例，提供了一种基于RDMA的组播方法，其可应用于分布式集合通信中的N个目的端之一。如图9所示，该组播方法可包括：步骤S910，基于第一UD类型QP接收组播报文，其中，所述组播报文携带分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；步骤S920，在确定所述组播报文携带的第一组播ID与自身的第二组播ID相同、并基于第一端点ID识别所述组播报文来自源端后，处理所述组播报文；步骤S930，当要向源端回复对所述组播报文的处理结果时，基于第二UD类型QP向源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及第二组播ID。

[0107] 可选地，如图9所示，在步骤S910之前，所述方法还包括：在目的端预先创建第一UD类型QP和第二UD类型QP(步骤S910p1)；以及，在目的端预先生成第二组播ID和第二虚拟端点ID(步骤S910p2)。

[0108] 可选地，步骤S910p2具体包括：根据目的端自身的排序值，按照预定的规则生成第二虚拟端点ID；以及，接收源端发送的组播IP地址，并根据所述组播IP地址、按照与源端生成第一组播ID相同的规则生成第二组播ID。这样，源端的第一组播ID以及各目的端的第二组播ID完全相同，可以用来标识由源端和所有目

的端构建的组播组内的报文。此外，源端的第一虚拟端点ID以及各目的端的第二虚拟端点ID相互不同，可以用来标识在该组播组内报文的具体来源。

[0109] 可选地，参考图5B所示，在基于第一UD类型QP接收携带数据报文标识以及待发送的数据包的组播数据报文(步骤⑤)之前，所述方法还包括：基于第一UD类型QP接收组播信息通告报文(步骤②)，其中，所述组播信息通告报文携带有信息通告标识以及与待传输数据相关的数据信息，所述数据信息包括待传输数据的数据量M、待传输数据被划分为K个数据块、每个数据块被进一步划分为s个数据包、以及各数据包的序列号，K为大于或等于1的整数，s为大于或等于1的整数；基于第二UD类型QP向源端发送信息通告确收报文(步骤③)，其中，所述信息通告确收报文携带有表示目的端接收到所述组播信息通告报文的确收标识。

[0110] 其中，如图5B所示，处理所述组播信息通告报文，可具体包括：根据所述组播信息通告报文携带的信息通告标识以及所述数据信息建立收包状态记录(步骤③)，其中，收包状态记录包括与每个数据包对应的状态标识，并且所述状态标识记录在与所述数据包的序列号对应的位置。此外，处理所述组播数据报文，可具体包括：根据所述组播数据报文携带的数据报文标识对所述组播数据报文携带的数据包进行接收处理，并更新目的端自身的收包状态记录中与所述数据包对应的状态标识(步骤⑥)。

[0111] 通过在各目的端上维护并更新收包状态记录，源端可以基于自身的发包记录以及各目的端上的收包状态记录，以数据块为单位进行批量的收包确认。具言之，如图5B(续)所示，在基于第一UD类型QP接收组播数据报文(步骤⑥)之后，所述方法还可包括：基于第一UD类型QP接收组播状态通告报文，其中，所述组播状态通告报文携带有状态通告标识(步骤⑦)；以及，根据所述组播状态通告报文生成状态响应报文，并基于第二UD类型QP向源端发送所述状态响应报文(步骤⑧)，其中，所述状态响应报文携带有所述目的端自身的所述收包状态记录。

[0112] 以上为本公开的一个或多个实施的方法，基于同样的思路，本公开还提供了相应的分布式集合通信的节点设备。参考图4所示，所述节点设备主要包括例如CPU、GPU的处理器以及能够支持RDMA的网卡。

[0113] 具体地，当该节点设备作为分布式集合通信中的源端410时，所述处理器被配置为：基于第一UD类型QP，经由所述网卡发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播数据报文携带的第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当基于第二UD类型QP经由所述网卡接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自目的端后，处理所述单播报文。

[0114] 此外，当该节点设备作为分布式集合通信中的目的端430-1、430-2、…、430-N时，所述处理器被配置为：基于第一UD类型QP，经由所述网卡接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；在确定所述组播报文携带的第一组播ID与所述目的端自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及，当要向源端回复对所述组播报文的处理结果时，基于第二UD类型QP，经由所述网卡向源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及第二组播ID，以使得源端在确定所述单播报文携带的第二组播ID与第一组播ID相同、并基于第二端点ID识别所述单播报文来自目的端后，处理所述单播报文。

[0115] 根据本公开的实施例，还提供了另一种分布式集合通信的节点设备。如图10所示，该节点设备1000主要包括收发模块1010和处理模块1020。

[0116] 具言之，当该节点设备1000作为分布式集合通信中的源端410时，所述处理模块1020被配置为：基于第一UD类型QP，经由所述收发模块1010发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播数据报文携带的第一组播ID与自身的第二组播ID相同、并基于第一端点ID识别所述组播报文来自源端

后，处理所述组播报文；以及，当基于第二UD类型QP经由所述收发模块1010接收到目的端发送的单播报文时，在确定所述单播报文携带的第二组播ID与第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自目的端后，处理所述单播报文。

[0117] 此外，当该节点设备1000作为分布式集合通信中的目的端430-1、430-2、…、430-N时，所述处理模块1020被配置为：基于第一UD类型QP，经由所述收发模块1010接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；在确定所述组播报文携带的第一组播ID与目的端自身的第二组播ID相同、并基于第一端点ID识别所述组播报文来自源端后，处理所述组播报文；以及，当要向源端回复对所述组播报文的处理结果时，基于第二UD类型QP，经由所述收发模块1010向源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及所述第二组播ID，以使得源端在确定所述单播报文携带的第二组播ID与第一组播ID相同、并基于第二端点ID识别所述单播报文来自目的端后，处理所述单播报文。

[0118] 示例性的，上述图4和图10中所示的节点设备可以是诸如台式电脑、笔记本电脑等终端设备，也可以是服务器或其他专用设备。上述图10所示的各个模块可以是硬件实现的模块，也可以是由软件程序实现的模块，还可以是软硬件相互配合以实现所述功能。这些模块仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分模块通过处理元件调用软件的形式实现，部分模块通过硬件的形式实现。例如，确定模块可以为单独设立的处理元件，也可以集成在上述装置的某一个芯片中实现，此外，也可以以程序代码的形式存储于上述装置的存储器中，由上述装置的某一个处理元件调用并执行以上确定模块的功能。其它模块的实现与之类似。此外这些模块全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件可以是一种集成电路，具有信号的处理能力。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

[0119] 例如，以上这些模块可以是被配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路(application specific integrated circuit，ASIC)，或，一个或多个微处理器(digital signal processor，DSP)，或，一个或者多个现场可编程门阵列(field programmable gate array，FPGA)等。再如，当以上某个模块通过处理元件调度程序代码的形式实现时，该处理元件可以是通用处理器，例如中央处理器(central processing unit，CPU)或其它可以调用程序代码的处理器。再如，这些模块可以集成在一起，以片上系统(system-on-a-chip，SOC)的形式实现。

[0120] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本公开实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘solid state disk(SSD))等。

[0121] 本公开还提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序可用于执行上述图8或图9提供的基于RDMA的组播方法。所述存储介质可以是非易失性存储介质。

[0122] 此外，图4所示的节点设备，在硬件层面，除了包括处理器和网卡，还可包括内存以及非易失性存储器，当然还可能包括其他业务所需要的硬件，如内部总线。处理器从非易失性存储器中读取对应的计算机程序到内存中然后运行，以实现上述图8或图9所述的基于RDMA的组播方法。

[0123] 当然，除了软件实现方式之外，本公开并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[0124] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0125] 本公开中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0126] 以上所述仅为本公开的实施例而已，并不用于限制本公开。对于本领域技术人员来说，本公开可以有各种更改和变化。凡在本公开的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本公开的权利要求范围之内。

权利要求书

[权利要求 1] 一种基于远程直接数据存取RDMA的组播方法，应用于分布式集合通信中的源端，所述方法包括：

基于第一不可靠数据报UD类型队列对QP发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播报文携带的所述第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及

当基于第二UD类型QP接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[权利要求 2] 根据权利要求1所述的方法，其特征在于，在基于所述第一UD类型QP发送组播报文之前，所述方法还包括：

生成组播互联网协议IP地址；

根据所述组播IP地址，按照预定的规则生成所述第一组播ID；以及向所述每个目的端发送所述组播IP地址，以使得每个目的端根据所述组播IP地址按照所述预定的规则生成所述第二组播ID。

[权利要求 3] 根据权利要求1或2所述的方法，其特征在于，在基于所述第一UD类型QP，发送组播报文之前，所述方法还包括：

随机生成所述第一虚拟端点ID。

[权利要求 4] 根据权利要求1至3中任一项所述的方法，其特征在于，所述基于第一UD类型QP发送组播报文，具体包括：

基于所述第一UD类型QP发送组播数据报文，其中，所述组播数据报文携带有数据报文标识以及待发送的数据包。

[权利要求 5] 根据权利要求4所述的方法，其特征在于，在所述基于所述第一UD类型QP发送组播数据报文之前，所述方法还包括：

基于所述第一UD类型QP发送组播信息通告报文，其中，所述组播信息通告报文携带有信息通告标识以及与待传输数据相关的数据信息；所述数据信息包括所述待传输数据的数据量M、所述待传输数据被划分为K个数据块、每个所述数据块被进一步划分为s个数据包、以及各所述数据包的序列号，K为大于或等于1的整数，s为大于或等于1的整数；

基于所述第二UD类型QP接收所述N个目的端返回的信息通告确收报文，其中，每个所述信息通告确收报文携带有表示所述目的端接收到所述组播信息通告报文的确收标识。

[权利要求 6] 根据权利要求5所述的方法，其特征在于，所述基于所述第二UD类型QP接收所述N个目的端返回的信息通告确收报文，具体包括：

在第一预定时长计时结束时判断已接收到的所述信息通告确收报文的个数是否小于N，其中，所述第一预定时长从发送所述组播信息通告报文时起算；

若前述判断的结果为是，则重新发送所述组播信息通告报文，并重新启动所述第一预定时长的计时，以返回判断已接收到的所述信息通告确收报文的个数是否小于N的步骤。

[权利要求 7] 根据权利要求5或6所述的方法，其特征在于，所述基于所述第一UD类型QP发送组播数据报文，具体包括：

针对所述K个数据块中每个数据块包括的待发送的每个数据包，基于所述第一UD类型QP发送携带该数据包的组播数据报文。

[权利要求 8] 根据权利要求7所述的方法，其特征在于，针对所述K个数据块中的每个数据块，在发送完该数据块包含的所有待发送数据包之后，所述方法还包括：

基于所述第一UD类型QP发送针对该数据块的组播状态通告报文，其中，所述组播状态通告报文携带有状态通告标识；

基于所述第二UD类型QP接收状态响应报文，其中，所述状态响应报文携带有所述目的端自身的收包状态记录；

依据该数据块下各数据包的发送记录以及接收到的所述状态响应报文中的收包状态记录，确定所述N个目的端是否接收到该数据块下的已发送数据包。

[权利要求 9] 根据权利要求8所述的方法，其特征在于，所述依据该数据块下各数据包的发送记录以及接收到的所述状态响应报文中的收包状态记录，确定所述N个目的端是否接收到该数据块下的已发送数据包，具体包括：

依据该数据块下各数据包的发送记录以及在第二预定时长内接收到的每个所述状态响应报文携带的收包状态记录，确定该数据块下的每个已发送的数据包是否被所有N个目的端接收，其中，所述第二预定时长从发送所述组播状态通告报文起算；

若确定存在未被所有N个目的端接收的数据包，则基于所述第一UD类型QP发送携带该数据包的组播数据报文，并重新启动所述第二预定时长的计时；

返回根据接收到的所述状态响应报文携带的收包状态记录确定已发送的数据包是否被所有N个目的端接收的处理，直到确定该数据块包括的所有数据包均已被所述N个目的端接收。

[权利要求 10] 根据权利要求9所述的方法，其特征在于，若在所述第二预定时长计时结束时未接收到任一所述状态响应报文，所述方法还包括：

重新发送所述组播状态通告报文，并

重新启动所述第二预定时长的计时。

[权利要求 11] 一种基于远程直接数据存取RDMA的组播方法，应用于分布式集合通信中的N个目的端中的一个所述目的端，N为大于等于2的整数，所述方法包括：

基于第一不可靠数据报UD类型队列对QP接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；

在确定所述组播报文携带的所述第一组播ID与所述目的端自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及

当要向所述源端回复对所述组播报文的处理结果时，基于第二UD类型QP向所述源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及所述第二组播ID，以使得所述源端在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[权利要求 12] 根据权利要求11所述的方法，其特征在于，在基于所述第一UD类型QP接收组播报文之前，所述方法还包括：

接收所述源端发送的组播互联网协议IP地址；

根据所述组播IP地址，按照与所述源端生成所述第一组播ID相同的规则生成所述第二组播ID。

[权利要求 13] 根据权利要求11或12所述的方法，其特征在于，在基于所述第一UD类型QP接收组播报文之前，所述方法还包括：

根据所述目的端自身的排序值，按照预定的规则生成所述第二虚拟端点ID。

[权利要求 14] 根据权利要求11至13中任一项所述的方法，其特征在于，所述基于所述第一UD类型QP接收组播报文，具体包括：

基于所述第一UD类型QP接收组播数据报文，其中，所述组播数据报文携带有数据报文标识以及待发送的数据包。

[权利要求 15] 根据权利要求14所述的方法，其特征在于，在所述基于所述第一UD类型QP接收组播数据报文之前，所述方法还包括：

基于所述第一UD类型QP接收组播信息通告报文，其中，所述组播信息通告报文携带有信息通告标识以及与待传输数据相关的数据信息；所述数据信息包括所述待传输数据的数据量M、所述待传输数据被划分为K个数据块、每个所述数据块被进一步划分为s个数据包、以及各所述数据包的序列号，K为大于或等于1的整数，s为大于或等于1的整数；

基于所述第二UD类型QP向所述源端发送信息通告确收报文，其中，所述信息通告确收报文携带有表示所述目的端接收到所述组播信息通告报文的确收标识。

[权利要求 16] 根据权利要求15所述的方法，其特征在于，

处理所述组播信息通告报文，包括：根据所述组播信息通告报文携带的所述信息通告标识以及所述数据信息建立收包状态记录，其中，所述收包状态记录包括与每个数据包对应的状态标识，并且所述状态标识记录在与所述数据包的序列号对应的位置；

处理所述组播数据报文，包括：根据所述组播数据报文携带的所述数据报文标识对所述组播数据报文携带的数据包进行接收处理，并更新所述目的端自身的所述收包状态记录中与所述数据包对应的状态标识。

[权利要求 17] 根据权利要求16所述的方法，其特征在于，在所述基于所述第一UD类型QP接收组播数据报文之后，所述方法还包括：

基于所述第一UD类型QP接收组播状态通告报文，其中，所述组播状态通告报文携带有状态通告标识；

根据所述组播状态通告报文生成状态响应报文，其中，所述状态响应报文携带有所述目的端自身的所述收包状态记录；以及

基于所述第二UD类型QP向所述源端发送所述状态响应报文。

[权利要求 18] 一种分布式集合通信的节点设备，作为所述分布式集合通信中的源端，包括处理器和支持远程直接数据存取RDMA的网卡，其特征在于，所述处理器被配置为：

基于第一不可靠数据报UD类型队列对QP，经由所述网卡发送组播报文，其中，所述组播报文携带第一虚拟端点ID以及第一组播ID，以使得所述组播报文经由网络交换机的组播功能转发至所述分布式集合通信中的N个目的端，N为大于或等于2的整数，并且每个目的端在确定所述组播数据报文携带的所述第一组播ID与自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及

当基于第二UD类型QP经由所述网卡接收到所述目的端发送的单播报文时，在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述单播报文携带的第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。

[权利要求 19] 一种分布式集合通信的节点设备，作为所述分布式集合通信中的N个目的端之一，包括处理器和支持远程直接数据存取RDMA的网卡，其特征在于，所述处理器被配置为：

基于第一不可靠数据报UD类型队列对QP，经由所述网卡接收组播报文，其中，所述组播报文携带所述分布式集合通信中的源端的第一虚拟端点ID以及第一组播ID；

在确定所述组播报文携带的所述第一组播ID与所述目的端自身的第二组播ID相同、并基于所述第一端点ID识别所述组播报文来自所述源端后，处理所述组播报文；以及

当要向所述源端回复对所述组播报文的处理结果时，基于第二UD类型QP，经由所述网卡向所述源端发送单播报文，其中，所述单播报文携带第二虚拟端点ID以及所述第二组播ID，以使得所述源端在确定所述单播报文携带的所述第二组播ID与所述第一组播ID相同、并基于所述第二端点ID识别所述单播报文来自所述目的端后，处理所述单播报文。